Instructional Science (2022) 50:361–390
https://doi.org/10.1007/s11251-022-09580-x

ORIGINAL RESEARCH

How many words are enough? Investigating the effect of different configurations of a software scaffold for formulating scientific hypotheses in inquiry-oriented contexts

Tasos Hovardas[1] · Zacharias Zacharia[1] · Nikoletta Xenofontos[1] · Ton de Jong[2]

Received: 16 November 2019 / Accepted: 12 February 2022 / Published online: 21 March 2022

Abstract

We extended research on scaffolds for formulating scientific hypotheses, namely the Hypothesis Scratchpad (HS), in the domain of relative density. The sample comprised of secondary school students who used three different configurations of the HS: Fully structured, containing all words needed to formulate a hypothesis in the domain of the study; partially structured, containing some words; unstructured, containing no words. We used a design with two different measures of student ability to formulate hypotheses (targeted skill): A global, domain-independent measure, and a domain-specific measure. Students used the HS in an intervention context, and then, in a novel context, addressing a transfer task. The fully and partially structured versions of the HS improved the global measure of the targeted skill, while the unstructured version, and to a lesser extent, the partially structured version, favored student performance as assessed by the domain-specific measure. The partially structured solution revealed strengths for both measures of the targeted skill (global and domain-specific), which may be attributed to its resemblance to completion problems (partially worked examples). The unstructured version of the HS seems to have promoted schema construction for students who revealed an improvement of advanced cognitive processes (thinking critically and creatively). We suggest that a comprehensive assessment of scaffolding student work when formulating hypotheses should incorporate both global and domain-specific measures and it should also involve transfer tasks.

Keywords Inquiry · Improving classroom science teaching · Pedagogical issues · Secondary education · Teaching/learning strategies

✉ Zacharias Zacharia
zach@ucy.ac.cy

[1] Research in Science and Technology Education Group, Department of Education, University of Cyprus, PO 20537, 1678 Nicosia, Cyprus

[2] Faculty of Behavioral, Management, and Social Sciences, University of Twente, GW/IST, PO Box 217, 7500 AE Enschede, The Netherlands

Introduction

Inquiry-based learning concentrates on knowledge and skill acquisition through self-regulated learning trajectories, which are taken over largely by learners themselves, and involve data selection, analysis and interpretation (de Jong & van Joolingen, 1998; de Jong, 2006a, b; Zacharia et al., 2015). Data processing should result in the discovery of relationships between the main variables in a domain (de Jong, 2006a; Bell et al., 2010). A related and recurrent difficulty encountered by students in inquiry-based learning settings is the formulation of testable hypotheses (de Jong & van Joolingen, 1998, van Joolingen et al., 2005). This requirement presupposes that students should be able to depict relations between the variables they identified and use correct syntax, namely, include a dependent and an independent variable in a hypothesized relation mediated by conditions, i.e., verbs detailing variable change: An "if–then" statement, with a conditional clause in the form of an "if-clause" and the consequence in the form of a following "then-clause" (van Joolingen & de Jong, 1991, 1993). Generating a hypothesis should be considered as a composite task, comprising the identification and selection of variables as an "entry" task, followed by interrelating variables, and finally, by restricting the range of the relations between variables by adding conditions (van Joolingen & de Jong., 1991).

Formulating hypotheses has a central position in inquiry-based learning because all subsequent student activities depend on their hypotheses, for instance, designing and performing an experiment during an investigation (de Jong, 2006b; Zacharia et al., 2015a, Efstathiou et al., 2018a; Klahr, 2005; Quintana et al., 2004). This central position, together with the difficulties encountered by students while formulating hypotheses, are the main reasons explaining why students need considerable guidance in this task (Chen et al., 2018; Oh, 2010). In computer-supported learning environments, the crucial importance of hypothesis formulation has been highlighted by the number of tools that have been developed to support this particular activity (van Joolingen et al., 2005a; Zacharia et al., 2015; Bell et al., 2010; Kim & Pedersen, 2011). A characteristic example of a software scaffold of this kind is the Hypothesis Scratchpad (HS) (van Joolingen & de Jong, 1991, 1993, 2003; de Jong, 2006b). The HS offers words to be considered by students when preparing their hypotheses. These words can be chosen by the teacher and they may refer to the variables involved or the nature of the relationship between variables. The HS provides dual support to students, since it assists them in variable selection, while it can also offer the proper syntax to portray a hypothesized relation between variables, e.g., in an "if …, then …" statement with conditions (Zacharia et al., 2015).

Previous research with the HS and the different configurations of the tool reflected several challenges in outlining good practice for providing proper guidance to students. In the first attempts to test the initial versions of the HS, it was found that a fully structured version of the tool, providing all words needed by students to formulate their hypotheses (i.e., variables, relations, and conditions), eventuated in better syntax of hypotheses than a partially structured or an unstructured version (van Joolingen & de Jong, 1991). However, the effect of full structure was not uniform across all parameters studied. For example, students working with the fully structured HS formulated a lower number of hypotheses and were less detailed in describing the relations between variables in their hypotheses as compared to students who worked with the partially structured or the unstructured HS (van Joolingen & de Jong, 1991). Overall, a major finding across studies was that students were overwhelmed by the complexity of the task (e.g., van Joolingen & de Jong, 1991, 1993, 1997). An alternative option was to increase structure further by providing pre-defined, complete, hypotheses and letting students elaborate on those (Gijlers & de Jong, 2009; de Jong, 2006b). Furnishing students with already generated hypotheses, however, may not allow them build a robust background schema on which to base their forthcoming experimentation (van Joolingen et al., 2005a). Despite the favorable outcomes of structuring (Gijlers & de Jong, 2009; de Jong, 2006a), a crucial consideration has always been that structuring learning activities beyond a certain level might not leave enough room for all the germane aspects and outcomes of inquiry that relate to challenging students (Gijlers & de Jong, 2005b).

The concern of the proper level of structure resembles the discussion of the optimal guidance vs. openness in inquiry-based learning environments (e.g., Arnold et al., 2014; Koksal & Berberoglou, 2014; see also Hmelo-Silver et al., 2007; Kirschner et al., 2006; Sweller et al., 2007). Optimizing inquiry learning approaches entails the need to simplify learning tasks by providing enough guidance, especially when students will encounter increased task complexity (e.g., van Joolingen & de Jong, 1997). However, rigid guidance might subtract freedom from students in enacting their explorations (e.g., Chang et al., 2008). In that case, guidance might compromise the opportunity for students to become autonomous in addressing novel learning contexts, and thereby, to secure learning gains in the long run. Analogous challenges have been also voiced in cognitive load research in the distinction between worked examples and partially worked examples (completion problems) (Baars et al., 2013; Paas, 1992; Sweller et al., 1998; Van Merriënboer, 1990, 1992), which echoes levels of varying structure offered to students. Partially worked examples have been suggested as superior to worked examples, because the latter provide a fully-fledged solution and may not let students engage deeply in the task at hand (e.g., Paas, 1992; Sweller et al., 1998; Van Merriënboer, 1990). In contrast, partially worked examples may stimulate a more comprehensive elaboration and deeper processing as long as they both provide guidance for initiating a task and necessitate a completion of the missing parts of the solution by the learner, which is expected to lead to a better quality of the solution schema (Baars et al., 2013).

Worked examples, completion problems (partially worked examples) and conventional problems (fully unstructured) may be re-conceptualized as cases along a gradient of decreasing structure (see in this regard Sweller et al., 1998) with important implications for learning and instruction. For instance, offloading should not be left to detract from challenging students to come up with a solution schema, which they could apply to new learning contexts (e.g., Paas, 1992; Van Merrienboer, 1992). Completion problems may be thought to satisfy both these needs, for example, decreasing extraneous load, as in the case of worked examples, and at the same time, facilitating schema construction (Sweller et al., 1998). Although previous research on the HS focused on structure as the type of support provided by the tool, it may have undervalued the germane aspects of student learning routes when the tool is not fully structured. In many cases, instruction should aim to engage students reflectively during learning trajectories, which requires a local increase in task complexity and letting students to take the initiative and resolve a situation on their own. These trade-offs between increasing and decreasing guidance and support may be exemplified in the learning task of formulating hypotheses by the number of words offered to students: The more words offered, the more structured the task would be. However, as the amount of words given to student increases, the less initiative and reflection are needed for screening and selecting the variables, relations, and conditions to include in their hypotheses. Another concern relates to student ability to successfully undertake learning tasks in different learning contexts. An arrangement of the HS may work well for a particular situation, but would it be equally effective in scaffolding the same learning task in a novel learning context?

Context and rationale for the present study

In this exploratory study, we aimed at extending research about the HS in the domain of relative density. In so doing, we involved secondary school students in order to explore varying guidance and support of this software scaffold. We conceptualized a gradient of decreasing structure (decreasing number of words offered to students to generate hypotheses) as analogous to the gradient formed by worked examples, partially worked examples (completion problems) and conventional problems (fully unstructured). Specifically, we explored learning outcomes for three configurations of the HS: Fully structured, including all words necessary to generate a hypothesis in the domain of the study; partially structured, including some words; unstructured, containing no words. The partially structured condition may be conceptualized as intermediate between the other two conditions, which is analogous to a completion problem. Presenting some words to students may catalyze the initiation of hypothesis generation (i.e., structuring function), but the rest of the words and their order would need to be produced by students themselves. To delve deeper into the effects of decreasing structure, we employed two different measures for the targeted skill (formulating hypotheses), a global, domain-independent measure, and a domain-specific measure. For the same reason, we also examined the influence of several process variables (i.e., time-on-task; products of learning activities) on the targeted skill. The inclusion of process variables allowed us to unravel the "black box" between frequently used pre-tests and post-tests and explore the effect of student learning routes on the improvement of the targeted skill, for instance, how students interacted with the learning environment. All these aspects are described in detail in the Methods section.

The design we followed together with the main variables we employed are depicted in Fig. 1. To investigate the effect of each condition of the HS on student ability to formulate hypotheses (targeted skill), we used two measures. The first was integrated in a pre- and post-test and was global in nature, i.e., did not explicitly address the domain where students worked (relative density). The second measure was directly linked to the domain and it was based on the classification of student hypotheses by means of a rubric in three categories: irrelevant or non-testable testaments; testable hypotheses; testable hypotheses with the interaction effect between the density of the object and fluid. We used a first learning context to familiarize students with the HS and another two learning contexts on relative density: An intervention context, where students experimented with a virtual lab; and a transfer context, where students undertook a transfer task. The choice of relative density for these two learning contexts allowed for examining student ability to go beyond testable hypotheses (i.e., beyond the selection of the right variables, relations and correct syntax) and detect the interaction effect between the density of object and fluid. This presupposed that students should have already constructed a basic schema to work with in the intervention context, which they could then apply to the transfer context.[1]

[1] To effectively manage a transfer task in a new learning setting, students need to identify both the surface features that may differ from the prior instructional context and the underlying core aspects shared by the two learning contexts (e.g., Schwartz et al., 2011; Shemwell et al., 2015). Learners would be expected to bypass surface features and apply the learned underlying core aspects that are shared between the previous learning setting and the novel setting (Barnett & Ceci, 2002; Belenky & Schalk, 2014; Kaminski et al., 2008).

The domain involved in the present study was sinking and floating. The primary difficulty that learners across age cohorts and educational levels need to overcome in this domain is their use of the spontaneous heuristic to concentrate on a single property (e.g., Potvin & Cyr, 2017). Most frequently, students concentrate on an object's mass to predict whether this object will sink or float in a fluid (Hsin & Wu, 2011; Loverude et al., 2003; Meindertsma et al., 2014). However, sinking or floating depend upon the density of the object and its relation to the density of the fluid. To arrive at density, learners must combine mass and volume. After learners have derived density from mass and volume, they must compare the density of the object to the density of the fluid (relative density) in order to make an informed judgment on whether an object will sink or float in a fluid. These two ratios (i.e., the ratio of mass to volume to determine density, and relative density), and more importantly, the interaction effect between the density of object and fluid, comprises the core underlying principle that operates throughout the domain, which needs to be applied in different learning contexts across that same domain. Adequate handling of a transfer challenge presupposes that a learner can distinguish between this core underlying principle, which is shared between learning contexts (i.e., the original instructional context and a new context in which the learner is requested to apply his/her knowledge and skills, i.e., transfer context), and other surface features that might vary between contexts. The learner must acknowledge the deep challenge common in the two learning contexts and bypass surface features that might differ and are unimportant for addressing new tasks. In that regard, the acknowledgment and use of the core underlying principle to be employed throughout the domain allows a learner to adequately handle transfer tasks (Chi & VanLehn, 2012; Schwartz et al., 2011; Shemwell et al., 2015).

Overall, we aimed at answering the following research questions:

1. What is the effect of the three configurations of the Hypothesis Scratchpad on cognitive processes and inquiry skills (including the global measure of the targeted skill)?
2. How do the three configurations of the Hypothesis Scratchpad differ in process variables?
3. What is the effect of process variables on the improvement of the global measure of the targeted skill, for each configuration of the Hypothesis Scratchpad?
4. How do the three configurations of the Hypothesis Scratchpad differ in their effect on the transfer task, as assessed by the domain-specific measure of the targeted skill?

Methods

Overview

Our study involved secondary school students in Cyprus, who worked in a computer-supported learning environment and used the HS to formulate their hypotheses. Three different classes were randomly assigned each to a different condition of the HS (Condition 1: Fully structured, all words; Condition 2: Partially structured, some words; Condition 3: Unstructured, no words). Each student worked individually with the same version of the HS in an initial learning context for familiarizing with the software scaffold ("weather" context; HS was used by students in a standalone fashion), an intervention context (here the HS was embedded in a learning activity sequence where students experimented in a virtual lab), and a context where students were asked to complete a transfer task ("submarine" context; HS was used in a standalone fashion). The intervention and transfer contexts were on the domain of relative density, where students had to detect the deep structure of the domain and incorporate it in their hypotheses, namely, the interaction effect between the density of object and fluid, which would determine if the object would sink or float. Student hypotheses in the intervention and transfer contexts were classified into categories by means of a rubric explicitly addressing the domain. This was the domain-specific measure we used to examine the targeted skill (i.e., formulating hypotheses). On top of this measure, students also completed a pre- and post-test, which included a global, domain-independent measure of the targeted skill, and a measure for identifying variables. The pre- and post-test also included two measures for cognitive processes related to transfer. Data collection evolved in two different but overlapping frames: The first focused on the global measure of the targeted skill. It begun with the pre-test, involved the familiarization and intervention contexts, and ended with the post-test. It also involved a number of process variables (time-on-task; products of learning activities) revealing aspects of student interaction with the learning environment in the intervention context (upper-left frame in Fig. 1). The second frame of data collection concentrated on the domain-specific measure of the targeted skill and it incorporated the intervention and transfer contexts (bottom-right frame in Fig. 1).

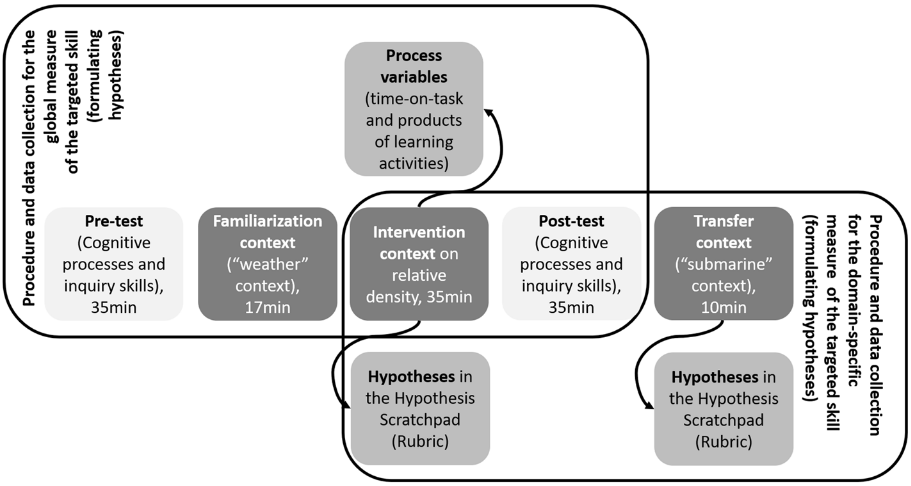
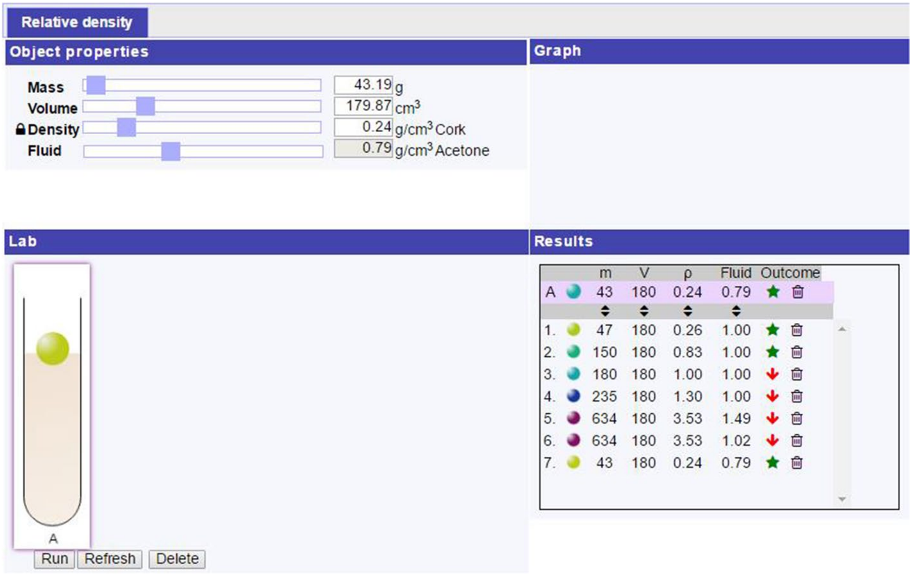

Procedure and data collection for the global measure of the targeted skill (formulating hypotheses)
Process variables (time-on-task and products of learning activities)
Pre-test (Cognitive processes and inquiry skills), 35min
Familiarization context ("weather" context), 17min
Intervention context on relative density, 35min
Post-test (Cognitive processes and inquiry skills), 35min
Transfer context ("submarine" context), 10min
Procedure and data collection for the domain-specific measure of the targeted skill (formulating hypotheses)
Hypotheses in the Hypothesis Scratchpad (Rubric)
Hypotheses in the Hypothesis Scratchpad (Rubric)

Fig. 1 The design of the study, learning contexts, and main measures/variables employed. Timeline can be followed from the left to the right, while the duration for the pre- and post-test as well as for the different learning contexts (familiarization; intervention; transfer) is given in each rectangle (light grey rectangles for the pre- and post-test; dark grey rectangles for the learning contexts). The upper-left box includes the procedure and data collection for the global measure of the targeted skill in a pre- and post-test arrangement, which involved a learning context to familiarize students with the Hypothesis Scratchpad and the intervention context. The targeted skill was measured in this case as an inquiry skill in the pre- and post-test (scale termed "Identifying and stating hypotheses" in TIPSII, see Online Appendix 2). Apart from the targeted skill, the pre- and post-test also included two cognitive processes (Online Appendix 1) and another one inquiry skill (Online Appendix 2). The bottom-right box includes the procedure and data collection for the domain-specific measure of the targeted skill in the intervention and transfer contexts. Here, the targeted skill was assessed by means of a rubric, which was used to categorize hypotheses formulated by students in the Hypothesis Scratchpad (see Fig. 3). Process variables (time-on-task and products of learning activities) were also examined in the intervention context (see Online Appendix 3 with details on process variables)

All aspects of our methodological approach are described in full detail in the following sub-sections.

Intervention context

Learning activities in the intervention context were undertaken online, in a computer-supported learning environment on relative density. The learning environment in the intervention context, henceforth called an Inquiry Learning Space (ILS), was developed by means of the Graasp authoring tool (de Jong et al., 2014, 2021) and followed the requirements outlined within the inquiry framework of Pedaste et al. (2015b). Learning activities were organized in separate phases, based on that framework. The first phase introduced students to the domain through a video that concentrated on the main variables they would encounter later on (*Orientation Phase*). The *Hypothesis Phase* came next, which included a virtual laboratory (Splash-Lab: "Splash: Virtual Buoyancy Laboratory"; http://www.golabz.eu/lab/splash-virtual-buoyancy-laboratory; Fig. 2). Students had the opportunity to explore the virtual laboratory and the variables to be manipulated (i.e., mass, volume and density of an object immersed in a fluid; density of the fluid; see Fig. 2; bars for manipulating variables shown in the top-left corner). After students had operationalized these variables, they could observe if the object sank or floated (see Fig. 2; animation available in the bottom-left corner). All values for all variables were given in a table (see Fig. 2; table in the bottom-right corner). A downward arrow in the table meant that the object sank, while a star indicated that the object floated. The last activity in the *Hypothesis Phase* was formulation of hypotheses, which was performed by means of the Hypothesis Scratchpad (see next sub-section in Methods).



Mass 43.19 g
Volume 179.87 cm^3
Density 0.24 g/cm^3 Cork
Fluid 0.79 g/cm^3 Acetone

	m	V	ρ	Fluid	Outcome
A ●	43	180	0.24	0.79	★
1. ●	47	180	0.26	1.00	★
2. ●	150	180	0.83	1.00	★
3. ●	180	180	1.00	1.00	↓
4. ●	235	180	1.30	1.00	↓
5. ●	634	180	3.53	1.49	↓
6. ●	634	180	3.53	1.02	↓
7. ●	43	180	0.24	0.79	★

Fig. 2 The Splash-Lab ("Splash: Virtual Buoyancy Laboratory"; http://www.golabz.eu/lab/splash-virtual-buoyancy-laboratory)

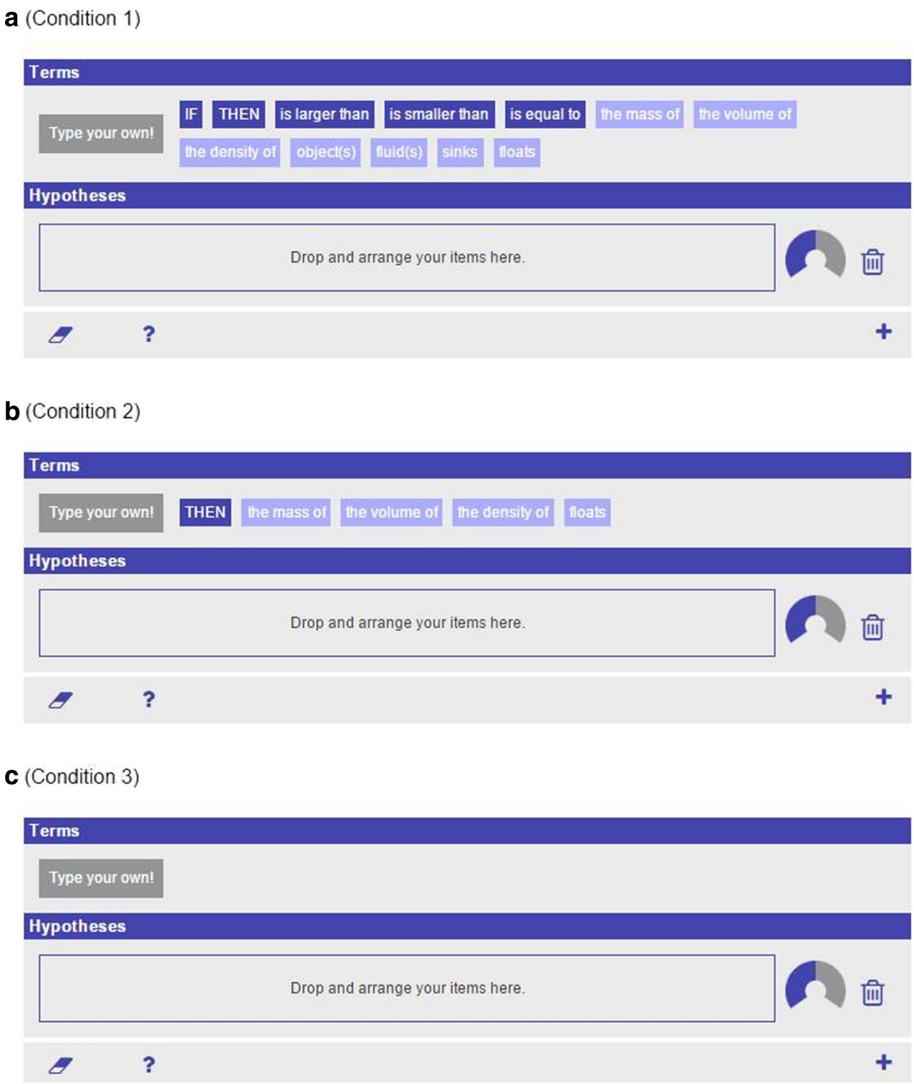


Fig. 3 The Hypothesis Scratchpad (https://www.golabz.eu/app/hypothesis-scratchpad); 3a corresponds to Condition 1, "all words"; 3b corresponds to Condition 2, "some words"; 3c corresponds to Condition 3, "no words"

Students then moved on to the *Investigation Phase*, where they conducted an experiment in the Splash-Lab to test their hypotheses. Students were prompted to keep notes of their observations using an observation tool. In the *Conclusion Phase*, students used their hypotheses and notes to reach a conclusion.

The Hypothesis Scratchpad

The HS (https://www.golabz.eu/app/hypothesis-scratchpad) was developed to support students in formulating hypotheses. It can include words to stimulate students to get started with the hypothesis formulation task (see Fig. 3a; upper part of the tool). Students can use these words or add their own, and then generate their hypothesis in the space provided (see Fig. 3a; lower part of the tool). If students wished to delete a word, they could use the eraser in the bottom-left corner of the tool. If they wished to delete an entire hypothesis, they could use the bin in the bottom-right corner. Three configurations of the HS were tested in the present study: The first version offered students all words needed to generate their hypotheses in the form of an "if…then" statement (Fig. 3a). The second configuration provided only a subset of words (see Fig. 3b; "then" shown in dark blue and a sub-set of variables shown in light blue). The words selected for this second configuration were the independent variables to be manipulated in the lab (i.e., "the mass of"; "the volume of"; "the density of"), the adverb "then", which opens up the conditional clause with the consequence, and "floats", which is one of the two different outcomes to be observed for the dependent variable. This selection of words required from students to: (1) select an independent variable; (2) describe whether this independent variable would be manipulated for the object or the fluid; (3) add the adverb for the conditional clause with the manipulation (i.e., "if"), (4) indicate a relation with a condition for linking the manipulation to the hypothesized outcome to be observed for the dependent variable (e.g., "is larger than"; "is smaller than"; "is equal to"); and to (5) determine if the outcome was the hypothesized one or if it should change to its rival outcome (i.e., "sinks"). All students in all conditions were notified that they could type in their own words and use them while formulating their hypotheses (see Fig. 3a–c; "Type your own box" upper left part of the tool). The third configuration of the HS included no provided words (Fig. 3c). In this case, the students had to type in themselves all of the words (variables, relations, conditions) needed for formulating a hypothesis. These three configurations corresponded to three conditions of varying support (i.e., all words: full structure; some words: partial structure; no words: no structure).

Participants

Participants were secondary school students who were guaranteed anonymity and participated in the research voluntarily after they themselves, and their parents, granted their informed consent. Students were notified that they had the option to withdraw at any stage from the study if they felt inclined to do so. No motive/reward was offered to students. The sample included 62 Greek Cypriot students from three different grade 9 classes in one school, who were of average ability in science and came from middle class families (mean age = 14.5 years; 28 boys, 45.2%; 34 girls, 54.8%). No participant had any prior experience with the HS. Each class was randomly assigned to one of the three conditions of the HS (Condition 1, "all words", 24 students; Condition 2, "some words", 18 students; Condition 3, "no words", 20 students). These numbers do not include two students in Condition 2 and one student in Condition 3 who had not provided a full series of data and were excluded from data analyses. Participants were not aware of condition assignments. There were no significant differences among conditions in terms of age or gender. Further, there was no significant difference among conditions on the pre-test (see last sub-section in Methods). There were no gender differences in cognitive processes or inquiry skills either before or after the instructional intervention.

Procedure

Implementation was carried out during regular school hours by one science teacher, who was trained to follow the same protocol. Students worked individually and they first completed the pre-test. Then, they got familiarized with the HS in the "weather" context (familiarization context), each one on his/her own computer in the Computer Lab of the school (17 min, on average). Then, each student accessed the ILS (intervention context) and worked individually; it took students about 35 min, on average, to go through the entire learning activity sequence concentrating on relative density. The only help students received from their teacher involved technical issues with regard to the use of the HS and the virtual laboratory. Whenever such technical issues occurred, they were resolved without causing any considerable delay in the completion of tasks. After exiting the ILS (intervention context), students completed the post-test. The last task involved using the HS in a stand-alone mode. Students were requested to formulate hypotheses in a new context (transfer context). Specifically, students were asked how a submarine can dive in the sea and re-surface, and whether they could think of any variables that might address this issue. Students were prompted to write down their hypotheses using the HS in the same version they had used it in the intervention context (10 min, on average). This last task was again performed by each student individually and it aimed at introducing a transfer challenge, namely, a task with different surface features (i.e., "submarine" instead of "object"; "sea" instead of "fluid"), but with the same underlying core principles related to the phenomenon under study (i.e., sinking or floating depends on relative density; the mass of the submarine may vary for the same volume when the tanks of the submarine are filled with water, and this causes the submarine to submerge or surface due to relative density). Three configurations of the HS were again prepared in this new, transfer context to align with the design in the intervention context and each student received the same tool configuration as in the intervention context. Throughout the procedure (familiarization context; intervention context; transfer context), each student worked individually. The teacher was instructed to resolve any issues with each student separately, and not to allow any interaction between students in the classroom. The intervention evolved as planned with no unexpected events.

Sources of data and coding

Three different data sources were used: (1) A pre- and post-test, which included a global measure of the targeted skill (formulating hypotheses); (2) data collected by means of computer screen capture software; and (3) the actual hypotheses that students generated in the intervention and transfer contexts, which were analyzed by means of a rubric to produce the domain-specific measure of the targeted skill.

Pre- and post-test

The pre- and post-test involved two instruments (cognitive processes; inquiry skills) both administered to students before and after the educational intervention. The instrument for cognitive processes was based on Bloom's (1956) taxonomy of educational objectives as it was revised by Anderson and Krathwohl (2001) and as it was further elaborated upon by de Jong (2014) and Zervas (2013). This instrument included items of two cognitive processes related to transfer, termed "Apply" and "Think critically and creatively" (Online

Appendix 1). The former measured student ability to apply knowledge already acquired to work through a new task, which was framed within a new learning context. The latter measured student ability to adapt acquired knowledge before addressing a novel context, which comprises screening background knowledge to select aspects which are relevant for addressing the novel context as well as combining these selected aspects to produce original meaning (see Efstathiou et al., 2018a for an analogous discussion of these cognitive processes). Both instruments were developed by a panel of four experts in science education and educational assessment and they were pilot tested with a sample of twenty students of the same age and ability as the sample recruited for the present study (the pilot sample was not included in the study sample). Minor edits were made after this pilot test, which verified the validity and reliability of the instrument. The instrument was developed and administered in Greek; it was translated by the third author to be included as Online Appendix 2 in this manuscript. We calculated inter-rater reliability (Cohen's Kappa) between two independent raters for the responses of the pilot sample in the open item in "Think critically and creatively", which amounted to 0.90.

The instrument for inquiry skills included items from the TIPSII instrument on "Identifying variables" (12 multiple-choice items, Online Appendix 2) and "Identifying and stating hypotheses" (9 multiple-choice items, Online Appendix 2) (see Burns et al., 1985, for a detailed description of all items and for the correct responses outlined for each item). This latter group of items on "Identifying and stating hypotheses" was used as the global measure of the targeted skill. The items were translated in Greek by the third author. A composite score was calculated for all cognitive processes and inquiry skills and rescaled to range between 0 (min) and 1 (max). In the post-test we used a different order of items in both instruments in order to mitigate the impact of the pre-test on the completion of the post-test. It took students 35 min, on average, to complete either the pre- or the post-test. By subtracting pre-test scores from post-test scores, we derived a measure of improvement in these measures following students' work in the intervention context.

Data collected by means of computer screen capture software

Using data collected through a computer screen capture software (River Past Screen Recorder Pro), we operationalized a series of process variables that reflected student interaction with the learning environment in the intervention context (Online Appendix 3). These included variables measuring time-on-task as well as variables associated with learning products, i.e., products created by students themselves, while undertaking learning activities. This is the complete list of process variables: (1) Overall time spent in the Hypothesis Phase (measured in seconds); (2) time spent in the Splash-Lab in the Hypothesis Phase (measured in seconds; included in overall time spent in the Hypothesis Phase); (3) time spent in the HS (measured in seconds; included in overall time spent in the Hypothesis Phase); (4) number of trials in the Splash-Lab in the Hypothesis Phase (count); (5) number of "smart" trials in the Splash-Lab in the Hypothesis Phase (count; "smart" trials in the Splash-Lab differed from other trials in the use of the "vary-one-variable-at-a-time" heuristic, where students kept either mass or volume of the object constant[2]); (6)

[2] Coding for the number of trials in the Splash-Lab versus "smart" trials in the lab (with the "vary-one-variable-at-a-time", VOTAT heuristic) as well as for the number of observations noted in the observation tool vs. "smart" observations (with a comparison between density of object and density of fluid) was performed by means of the computer screen capture software (River Past Screen Recorder Pro) and did not necessitate any control for inter-rater reliability.

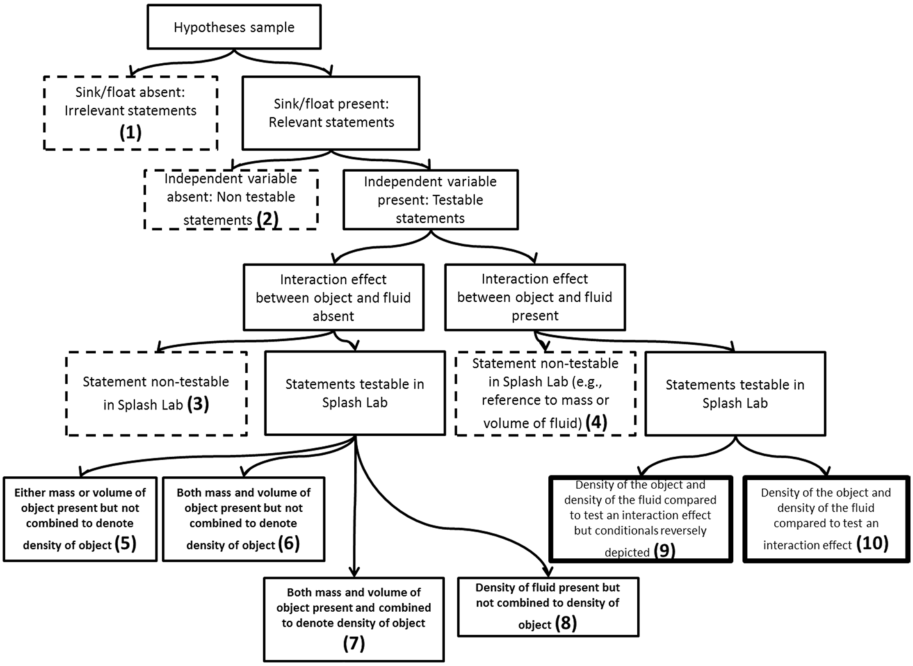

Hypotheses sample
Sink/float absent: Irrelevant statements (1)
Sink/float present: Relevant statements
Independent variable absent: Non testable statements (2)
Independent variable present: Testable statements
Interaction effect between object and fluid absent
Interaction effect between object and fluid present
Statement non-testable in Splash Lab (3)
Statements testable in Splash Lab
Statement non-testable in Splash Lab (e.g., reference to mass or volume of fluid) (4)
Statements testable in Splash Lab
Either mass or volume of object present but not combined to denote density of object (5)
Both mass and volume of object present but not combined to denote density of object (6)
Density of the object and density of the fluid compared to test an interaction effect but conditionals reversely depicted (9)
Density of the object and density of the fluid compared to test an interaction effect (10)
Both mass and volume of object present and combined to denote density of object (7)
Density of fluid present but not combined to density of object (8)

Fig. 4 Rubric employed to categorize student hypotheses in the intervention and transfer contexts. Rectangles with dashed lines depict irrelevant or non-testable statements (1, 2, 3, 4). Rectangles with continuous lines stand for testable statements without an interaction effect between the density of object and fluid (5, 6, 7, 8), while rectangles with bold lines correspond to testable statements with the interaction effect (9, 10)

overall time spent in the Investigation Phase (measured in seconds); (7) time spent in the Splash-Lab in the Investigation Phase (measured in seconds; included in overall time spent in the Investigation Phase); (8) number of trials in the Splash-Lab in the Investigation Phase (count); (9) number of "smart" trials in the Splash-Lab in the Investigation Phase (count); (10) number of observations noted in the observation tool in the Investigation Phase and after students had used the Splash-Lab (count); (11) number of "smart" observations (count; "smart" observations differed from other observations in that they included a comparison of the density of the object with the density of the fluid); (12) time spent in the Conclusion Phase (measured in seconds). Time spent in the Orientation Phase was not included in the analysis since this was equal to the duration of the video and did not differ between students.

Overall, we included six process variables measuring time-on-task (1, 2, 3, 6, 7, and 12) and another six process variables associated with learning products created by students during the learning activity sequence (4, 5, 8, 9, 10, and 11). Time-on-task was calculated as time devoted by students to working in an entire phase (1, 6, 12) or time devoted to working in either the HS (3) or the Splash-Lab (2, 7). To calculate time-on-task we employed a fine-grained approach, and distinguished on-task from off-task actions while observing screen-capture data. For instance, we isolated time spent on other websites than the one hosting the ILS (intervention context) used in the implementation (see, for example, Xenofontos et al., 2020; Cohen et al., 2007). Time spent on the former was not included in time-on-task.

Student hypotheses

We developed a rubric to classify hypotheses that students formulated in the HS both in the intervention and transfer contexts. This was used to produce the domain-specific measure of the targeted skill. For students who formulated more than one hypothesis in either context, the hypothesis with the highest score was selected for data analyses. Hypotheses were first assigned to ten different categories (Fig. 4) and then re-assigned to three broader categories: (1) Irrelevant statements or statements that could not be tested in the SplashLab; (2) testable statements without interaction effect between the density of the object and fluid; and (3) testable statements with interaction effect between density of object and fluid. Two raters, independently, coded hypotheses; their inter-rater reliability (Cohen's Kappa), calculated for the entire set of the initial ten categories in the rubric, was 0.82. The mismatches were assigned after a final discussion between coders.

Statistical analyses

Since our data had non-parametric distribution, data analyses involved non-parametric tests. Kruskal–Wallis and Mann–Whitney tests were used to investigate if there were significant differences among conditions in cognitive processes and inquiry skills, including the global measure of the targeted skill (formulating hypotheses), and process variables. Wilcoxon signed rank tests were conducted to examine temporal trends for each condition (differences between the pre-test and the post-test) in cognitive processes and inquiry skills, including the global measure of the targeted skill. We used the Bonferroni correction in all the statistical results we present for these tests. Tree modeling was employed to examine the effect of process variables on improvement in the global measure of the targeted skill. To investigate change in the domain-specific measure of the targeted skill, we concentrated on the hypotheses generated by students in the intervention and transfer contexts. Another tree model was computed to examine the effect of cognitive skills, inquiry skills and process variables on the domain-specific measure of the targeted skill. Online Appendix 4 presents measures, type of measure, instrument/source of data, and data analyses in which measures were used).

Results

Preliminary analysis: How did the two measures we used for assessing the targeted skill (global measure; domain-specific measure) interrelate?

We begin the Results section with a preliminary analysis to examine if the two different measures of the targeted skill (formulating hypotheses) were interrelated (global measure; domain-specific measure).[3] For the total sample, pre-test scores for the global measure of

[3] The global and domain-specific measures of the targeted skill (hypothesis formulation) should be related somehow, since we should have expected that a student scoring high in the global (domain-independent) measure should also be capable of addressing effectively the domain-specific task as assessed by means of the rubric. This is what we examined in "Preliminary analysis" through non-parametric analyses (global measure treated as scale variable; domain-specific measure treated as nominal variable). A possible relation between the two measures should not lead us to collapse the two measures into one, however, since the first, global measure (scale variable) would still denote the targeted skill in a context-independent manner, while the second, domain-specific measure (nominal variable) would be confined within the frame of relative density (domain of the present study). In this domain, formulating testable hypotheses is not enough, since students also need to identify and incorporate in their hypotheses the interaction effect between the density of object and fluid.

the targeted skill (as assessed by means of the TIPSII items) differed significantly for students whose hypotheses generated in the intervention context fell into the different categories as assigned by the rubric (domain-specific measure of the targeted skill) (Kruskal–Wallis $\chi^2 = 18.98$, $p < 0.001$). Students who formulated testable statements with the interaction effect between the density of the object and fluid in the intervention context (mean value for the global measure = 0.47) outperformed those who formulated testable statements without the interaction effect (mean value for the global measure = 0.32; Mann–Whitney $Z = -3.21$, $p < 0.001$) or those who formulated irrelevant or non-testable statements (mean value for the global measure = 0.22; Mann–Whitney $Z = -3.92$, $p < 0.001$). This meant that the two different measures we used to assess the targeted skill (global measure as assessed by means of the TIPSII items used in pre- and post-tests; domain-specific measure assessed by means of the rubric with categories of hypotheses capturing the interaction effect between the density of the object and fluid, which reflects the deep underlying principle of the domain of relative density) were aligned.

We performed another check to cross-validate the alignment between the two measures of the targeted skill. This involved the transition from the intervention context to the transfer context ("submarine"). We tracked differences among conditions in the domain-specific measure of the targeted skill, namely, in the classification of student hypotheses using the three categories of the rubric. A student's hypothesis could be assigned to the same category in the intervention and transfer contexts, could progress and move upwards (e.g., move from formulating irrelevant or non-testable statements to formulating testable statements without or with the interaction effect) or could regress and move downwards in the classification (e.g., move from formulating testable statements with the interaction effect to statements without this effect). Progress was not possible for the upper level of the classification (testable statements with the interaction effect between the density of the object and fluid; only skill maintenance was possible for this category), while regress was not possible for the lower level of the classification (irrelevant or non-testable statements).

We found that improvement in the global measure of the targeted skill, calculated as difference between post-test and pre-test scores on TIPSII items, differed significantly among students, who showed progress in the domain-specific measure of the targeted skill, that is, in the transition from the intervention context to the transfer context, in terms of the level of hypotheses generated, as compared to those who either remained in the same category or regressed (Kruskal–Wallis $\chi^2 = 6.71$, $p < 0.05$). Specifically, the mean gain score for TIPSII items for students who progressed was 0.20, which differed significantly from the mean value for students who remained in the same category (mean value for TIPSII items = 0.07; Mann–Whitney $Z = -2.27$, $p < 0.05$) or students who regressed (mean value for TIPSII items = 0.09; Mann–Whitney $Z = -2.39$, $p < 0.05$). This finding indicated, once again, that improvement in the global measure of the targeted skill aligned with improvement in the domain-specific measure of the targeted skill.

Table 1 Average values for cognitive processes and inquiry skills across conditions

	Condition 1 (all words; n = 24)	Condition 2 (some words; n = 18)	Condition 3 (no words; n = 20)	Kruskal–Wallis test (χ^2)
"Apply" (Cognitive process)				
Pre-test	0.36 (0.34)	0.54 (0.38)	0.37 (0.30)	2.87^{ns}
Post-test	0.49 (0.39)	0.61 (0.33)	0.60 (0.34)	1.30^{ns}
Wilcoxon signed ranks test (Z)	-1.82^{ns}	-0.93^{ns}	-2.86**	
Improvement	0.13 (0.35)	0.07 (0.42)	0.23 (0.29)	3.19^{ns}
"Think critically and creatively" (Cognitive process)				
Pre-test	0.22 (0.25)	0.39 (0.29)	0.23 (0.23)	4.44^{ns}
Post-test	0.39 (0.29)	0.42 (0.34)	0.38 (0.30)	0.18^{ns}
Wilcoxon signed ranks test (Z)	-3.02**	-0.30^{ns}	-2.21^{ns}	
Improvement	0.17 (0.24)	0.03 (0.26)	0.15 (0.25)	3.43^{ns}
"Identifying variables" (Inquiry skill)				
Pre-test	0.35 (0.18)	0.41 (0.19)	0.38 (0.20)	1.98^{ns}
Post-test	0.57 (0.18)	0.51 (0.20)	0.50 (0.19)	1.84^{ns}
Wilcoxon signed ranks test (Z)	-4.30***	-2.24^{ns}	-3.17**	
Improvement	0.22 (0.12)	0.10 (0.14)	0.12 (0.11)	9.20^{ns}
"Identifying and stating hypothesis" (Inquiry skill; global measure of the targeted skill)				
Pre-test	0.35 (0.15)	0.33 (0.19)	0.40 (0.22)	0.83^{ns}
Post-test	0.50 (0.16)	0.48 (0.19)	0.42 (0.21)	2.65^{ns}
Wilcoxon signed ranks test (Z)	-3.67***	-3.18**	-0.65^{ns}	
Improvement	0.15 (0.14)	0.15 (0.14)	0.02 (0.10)	11.91**

Items for cognitive processes and inquiry skills are presented in Online Appendices 1 and 2, respectively; "improvement" was calculated by subtracting pre-test scores from post-test-scores; average values presented were recalculated to range between 0 and 1; standard deviations are given in parentheses; *ns* non-significant; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; the Bonferroni correction was used for multiple comparisons

Research question 1: What is the effect of the three configurations of the Hypothesis Scratchpad on cognitive processes and inquiry skills (including the global measure of the targeted skill)?

Table 1 presents non-parametric tests for all cognitive processes and inquiry skills examined including the global measure of the targeted skill ("Identifying and testing hypotheses"). Across all measures, "improvement" was calculated by subtracting pre-test scores from post-test-scores. The only significant difference between conditions was observed for improvement in the global measure of the targeted skill (Kruskal–Wallis $\chi^2 = 11.91$, $p < 0.01$). Improvement was higher in Condition 1 (fully structured; all words) and Condition 2 (partially structured; some words) compared to Condition 3 (unstructured; no words) (Mann–Whitney $Z = -2.99$, $p < 0.01$, and Mann–Whitney $Z = -3.01$, $p < 0.01$, respectively). Wilcoxon tests performed for each condition separately showed that both Conditions 1 and 2 progressed in the global measure of the targeted skill (Wilcoxon $Z = -3.67$, $p < 0.001$ for Condition 1; Wilcoxon $Z = -3.18$, $p < 0.01$ for Condition 2), but Condition 3 did not. Although all conditions showed higher post-test scores than pre-test scores for all other measures in Table 1, significant trends were only revealed for "Think critically and creatively" in Condition 1 (Wilcoxon $Z = -3.02$, $p < 0.01$), "Apply" for Condition 3

Table 2 Average values for process variables in the intervention context across conditions

	Condition 1 (all words; n = 24)	Condition 2 (some words; n = 18)	Condition 3 (no words; n = 20)	Kruskal–Wallis χ^2
Time spent in the hypothesis phase (seconds)	643.13 (210.16)	738.61 (196.84)	779.35 (283.17)	4.60^{ns}
Time spent in the Splash Lab in the hypothesis phase (seconds)	206.38 (107.29)	277.06 (135.64)	226.30 (64.63)	4.24^{ns}
Time spent in the hypothesis scratchpad (seconds)	293.33 (158.61)	330.33 (126.93)	332.35 (155.24)	1.95^{ns}
Number of trials in the Splash Lab in the hypothesis phase (count)	7.63 (6.52)	9.89 (7.00)	12.40 (10.41)	4.17^{ns}
Number of “smart” trials in the Splash Lab in the hypothesis phase (count)	2.21 (2.84)	5.00 (4.31)	6.40 (7.47)	7.44^{ns}
Time spent in the Investigation Phase (seconds)	599.17 (248.09)	469.11 (143.83)	454.10 (137.97)	6.08^{ns}
Time spent in the Splash Lab in the investigation phase (seconds)	152.71 (103.32)	116.67 (73.85)	165.65 (88.84)	3.19^{ns}
Number of trials in the Splash Lab in the investigation phase (count)	10.17 (9.44)	11.22 (8.74)	10.10 (6.10)	0.58^{ns}
Number of “smart” trials in the Splash Lab in the investigation phase (count)	5.13 (6.79)	7.33 (7.94)	5.65 (4.55)	1.63^{ns}
Number of observations (count)	1.67 (0.92)	1.78 (0.55)	2.30 (1.30)	6.33^{ns}
Number of “smart” observations (count)	0.54 (0.78)	0.94 (0.87)	1.50 (0.95)	11.02^{**}
Time spent in the conclusion phase (seconds)	303.21 (132.66)	253.00 (185.40)	313.30 (181.11)	1.10^{ns}

“Smart” trials in the Splash Lab differed from other trials in the use of the “vary-one-variable-at-a-time” heuristic, where students kept either mass or volume of the object constant; “smart” observations differed from other observations in that they included a comparison of the density of the object with the density of the fluid; standard deviations are given in parentheses; *ns* non-significant; $^*p < 0.05$; $^{**}p < 0.01$; the Bonferroni correction method was used for multiple comparisons

(Wilcoxon $Z = -2.86$, $p < 0.01$) and "Identifying variables" for Conditions 1 and 3 (Wilcoxon $Z = -4.30$, $p < 0.001$, and Wilcoxon $Z = -3.17$, $p < 0.01$, respectively).

Research question 2: How do the three configurations of the Hypothesis Scratchpad differ in process variables?

Mean scores for process variables across conditions in the intervention context are presented in Table 2. The only significant difference was for number of "smart" observations recorded, which differed from other observations in that they included a comparison of the density of the object with the density of the fluid (Kruskal–Wallis $\chi^2 = 11.02$, $p < 0.01$), and where Condition 3 (unstructured; no words) showed the highest mean score. Although there was a general trend with Condition 3 delivering relatively more products for learning activities in the Hypothesis Phase (trials and "smart" trials in the virtual laboratory) and the Investigation Phase (e.g., observations) as compared to the other conditions, these latter differences were not significant.

Research question 3: What is the effect of process variables on the improvement of the global measure of the targeted skill, for each configuration of the Hypothesis Scratchpad?

Tree modelling was employed to examine the effect of process variables on improvement in the global measure of the targeted skill across conditions, as calculated by gain scores from pre-test to post-test. Figure 5 depicts the tree for Condition 1 (fully structured; all words). At each split, process variables are shown together with values partitioning the student sub-sample at each branch (i.e., left and right branches). Each node shows the mean value and standard deviation of the gain, number of students (n) and percentage of the student sample. We can read the tree by moving from the top downwards, up to each end node. In the first split, the number of "smart" trials in the Splash-Lab in the Investigation Phase meant improvement for the majority of students in Condition 1 on the right branch (Node 2, n = 17), while students who failed to perform more than one "smart" trial were allocated to the left branch of the tree (Node 1, n = 7). Following the right branch of the tree to the next split, we can observe that there was a threshold (579.5 s), after which time spent in the Investigation Phase did not favor improvement: Students who spent less than the threshold (Node 5, n = 7) showed higher improvement than those who spent more than this threshold time (Node 6, n = 10). Taken together, these findings indicate that more than one "smart" trial in the Splash-Lab in the Investigation Phase and less than the threshold time in the Investigation Phase (579.5 s) led to maximum improvement in the global measure of the targeted skill (Node 5).

Figure 6 displays the tree for Condition 2 (partially structured; some words). In this case, in the first split, improvement of the global measure of the targeted skill increased with number of observations for the majority of students (Node 2, n = 13). Figure 7 presents the tree for Condition 3 (unstructured; no words). Here improvement of the global measure of the targeted skill was favored by number of "smart" trials in the Splash-Lab in the Investigation Phase (first split, right branch, Node 2, n = 18) and time spent in the Splash-Lab in the Hypothesis Phase (second split, right branch, Node 4, n = 14). In the next split on the right half of the tree, there was a threshold related to student usage of the

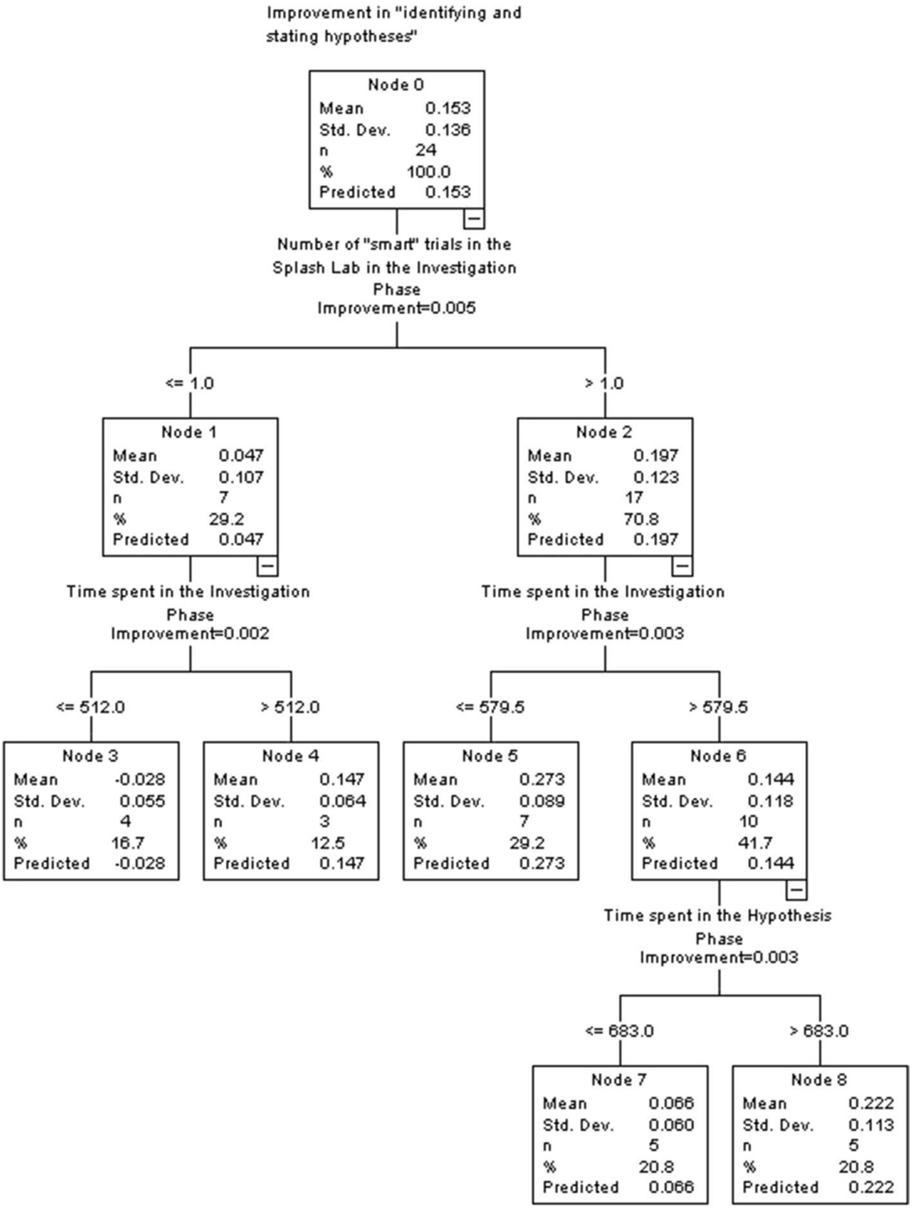
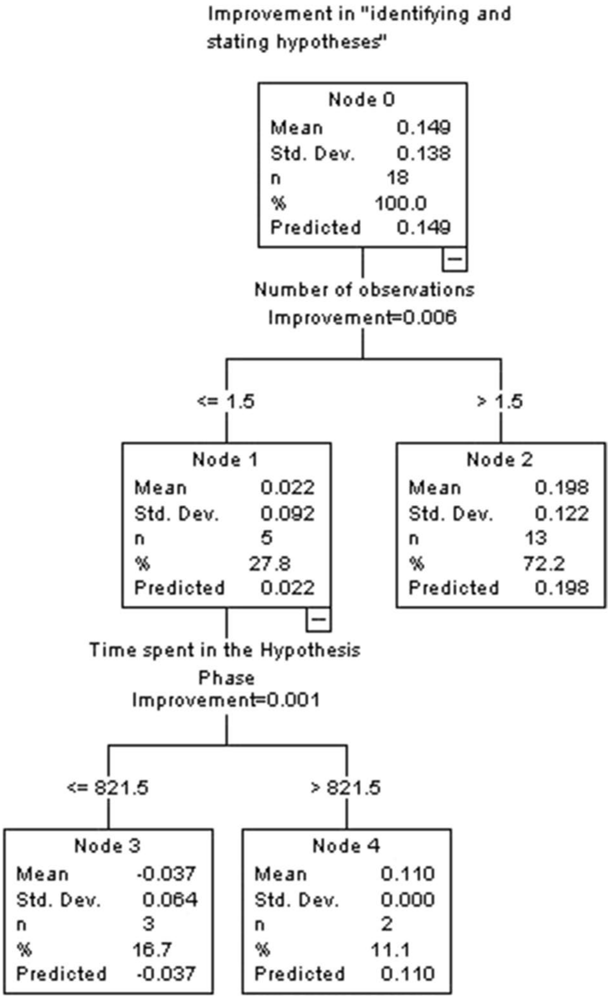


Fig. 5 Tree model for improvement in "Identifying and stating hypotheses" (global measure of the targeted skill) in Condition 1 (all words; n = 24). Process variables are shown at each split together with thresholds for partitioning the student sub-sample at each branch (i.e., left and right branches). Each node shows the mean value and standard deviation of improvement (gain score) in the targeted skill, number of students (n) and percentage of the student sample. A negative mean denotes that the post-test score for the targeted skill was lower than the pre-test score. Total variance explained by the tree = 72.97%

Fig. 6 Tree model for improvement in "Identifying and stating hypotheses" (global measure of the targeted skill) in Condition 2 (some words; n = 18). Process variables are shown at each split together with thresholds for partitioning the student sub-sample at each branch (i.e., left and right branches). Each node shows the mean value and standard deviation of improvement (gain score) in the targeted skill, number of students (n) and percentage of the student sample. A negative mean denotes that the post-test score for the targeted skill was lower than the pre-test score. Total variance explained by the tree = 94.75%



Splash-Lab in the Investigation Phase, after which improvement no longer advanced (Node 7, 174.5 s).

For all trees (Figs. 5, 6, 7), there were two main findings that need to be highlighted. First, learning products based on the interaction of students with the learning environment (number of "smart" trials in the Splash-Lab in the Investigation Phase for Conditions 1 and 3; number of observations for Condition 2) were featured in the first splits, meaning that these process variables were most decisive for the improvement in the global measure of the targeted skill across all conditions. Second, there were some thresholds for overall time spent in the Investigation Phase (Condition 1) or in the Splash-Lab in the Investigation Phase (Condition 3), after which improvement was no longer facilitated. We should note that no such threshold was revealed by our trees for any dimension of time-on-task in the Hypothesis Phase.

Fig. 7 Tree model for improvement in "Identifying and stating hypotheses" (global measure of the targeted ▶ skill) in Condition 3 (no words; n = 20). Process variables are shown at each split together with thresholds for partitioning the student sub-sample at each branch (i.e., left and right branches). Each node shows the mean value and standard deviation of improvement (gain score) in the targeted skill, number of students (n) and percentage of the student sample. A negative mean denotes that the post-test score for the targeted skill was lower than the pre-test score. Total variance explained by the tree = 89.59%

Research question 4: How do the three configurations of the Hypothesis Scratchpad differ in their effect on the transfer task, as assessed by the domain-specific measure of the targeted skill?

Table 3 presents the distribution according to assigned categories of the hypotheses formulated by students in the intervention context and transfer context ("submarine") across conditions (domain-specific measure of the targeted skill). A likelihood ratio chi-square test revealed a significant result in the transfer context ($\chi^2 = 26.39$, $p < 0.001$; Cramér*'s* $V = 0.42$, $p < 0.001$), with no student in Condition 1 (fully structured; all words) managing to formulate a testable statement with interaction effect between the density of object and fluid. Specifically, none of the seven students who had identified such an interaction effect in the intervention context managed to do so in the transfer context and no student with testable statements without an interaction effect progressed in the transfer context to including this effect. In the transfer context, Condition 2 (partially structured; some words) presented an accumulation of students in the middle category (testable statements without interaction effect between the density of object and fluid), while three students only included the interaction effect in their hypotheses. In Condition 3 (unstructured; no words) the student sub-sample was split into the extreme categories (i.e., irrelevant or non-testable statements and testable statements with interaction effect between the density of object and fluid). When taking into account student progress in the categories of the rubric in moving from the intervention context to the transfer context, we found that student capacity to progress along the categories of the rubric was marginally higher in Condition 2, with more than one-fourth of the students in that condition following that trend (27.28% in Condition 2, as compared to 20.00% in Condition 3 and 16.67% in Condition 1).

To study the differences between conditions in transfer further, we computed another tree model with the categories of hypotheses in the transfer context as dependent variable (domain-specific measure of the targeted skill) (Fig. 8). In this case, we included among independent variables the configuration of the HS (Conditions 1–3), cognitive processes ("Apply"; "Think critically and creatively"), the inquiry skill termed "Identifying variables", and process variables (time-on-task variables; products of learning activities). The configuration of the HS partitioned the sample in two different branches (first split), with Conditions 1 (fully structured; all words) and 2 (partially structured; some words) being arranged on the left branch of the tree and Condition 3 (unstructured; no words) on the right branch. A higher score than the threshold set by the tree model (0.17) for "Identifying variables" (inquiry skill) already in the pre-test increased the odds for students in Conditions 1 and 2 to generate testable hypotheses in the transfer context (left branch of the tree; second split). Students in Condition 3 were more inclined to include in their hypotheses the interaction effect between the density of object and fluid if they had a score over the threshold set by the tree model (0.17) for the cognitive process "Think critically and creatively" in the post-test (right branch of the tree; second split). Overall, student achievement in the transfer task (domain-specific measure of the targeted skill) was mediated for Conditions 1 and 2 by the entry inquiry skill "Identifying hypotheses" already from the pre-test. In

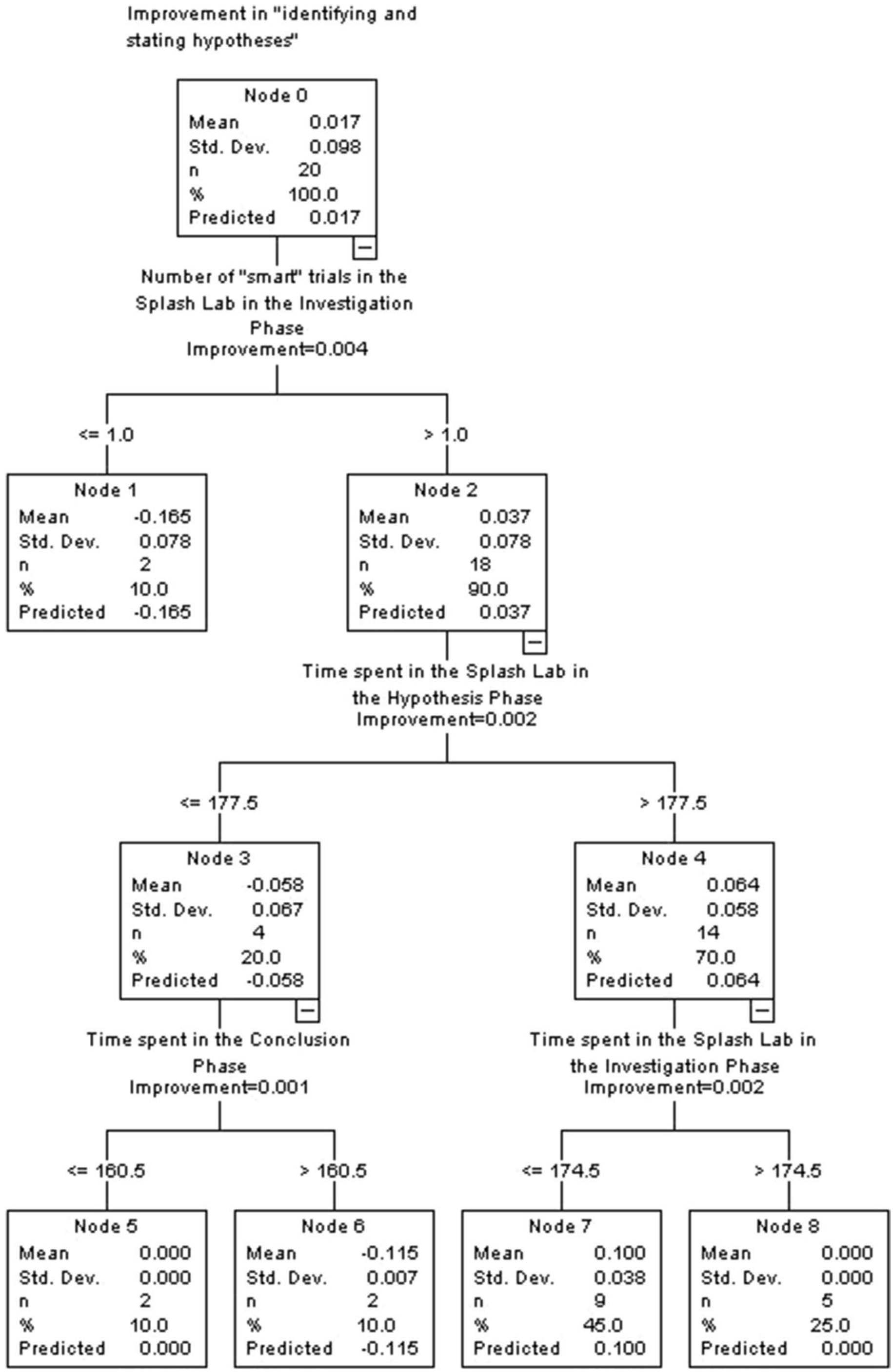

Improvement in "identifying and stating hypotheses"
Node 0
Mean 0.017
Std. Dev. 0.098
n 20
% 100.0
Predicted 0.017
Number of "smart" trials in the Splash Lab in the Investigation Phase
Improvement=0.004
<= 1.0
> 1.0
Node 1
Mean -0.165
Std. Dev. 0.078
n 2
% 10.0
Predicted -0.165
Node 2
Mean 0.037
Std. Dev. 0.078
n 18
% 90.0
Predicted 0.037
Time spent in the Splash Lab in the Hypothesis Phase
Improvement=0.002
<= 177.5
> 177.5
Node 3
Mean -0.058
Std. Dev. 0.067
n 4
% 20.0
Predicted -0.058
Node 4
Mean 0.064
Std. Dev. 0.058
n 14
% 70.0
Predicted 0.064
Time spent in the Conclusion Phase
Improvement=0.001
Time spent in the Splash Lab in the Investigation Phase
Improvement=0.002
<= 160.5
> 160.5
<= 174.5
> 174.5
Node 5
Mean 0.000
Std. Dev. 0.000
n 2
% 10.0
Predicted 0.000
Node 6
Mean -0.115
Std. Dev. 0.007
n 2
% 10.0
Predicted -0.115
Node 7
Mean 0.100
Std. Dev. 0.038
n 9
% 45.0
Predicted 0.100
Node 8
Mean 0.000
Std. Dev. 0.000
n 5
% 25.0
Predicted 0.000

Table 3 Distribution of categories of hypotheses in the intervention context on relative density and the transfer context ("submarine" context) across conditions

	Intervention context on relative density (% of students)	Transfer context ("submarine" context) (% of students)
Condition 1 (all words; n = 24)		
Irrelevant or non-testable statements	6 (25.0)	6 (25.0)
Testable statements without interaction effect between object and fluid	11 (45.8)	18 (75.0)
Testable statements with interaction effect between object and fluid	7 (29.2)	0 (0.0)
Condition 2 (some words; n = 18)		
Irrelevant or non-testable statements	3 (16.7)	4 (22.2)
Testable statements without interaction effect between object and fluid	8 (44.4)	11 (61.1)
Testable statements with interaction effect between object and fluid	7 (38.9)	3 (16.7)
Condition 3 (no words; n = 20)		
Irrelevant or non-testable statements	1 (5.0)	7 (35.5)
Testable statements without interaction effect between object and fluid	10 (50.0)	3 (15.0)
Testable statements with interaction effect between object and fluid	9 (45.0)	10 (50.0)

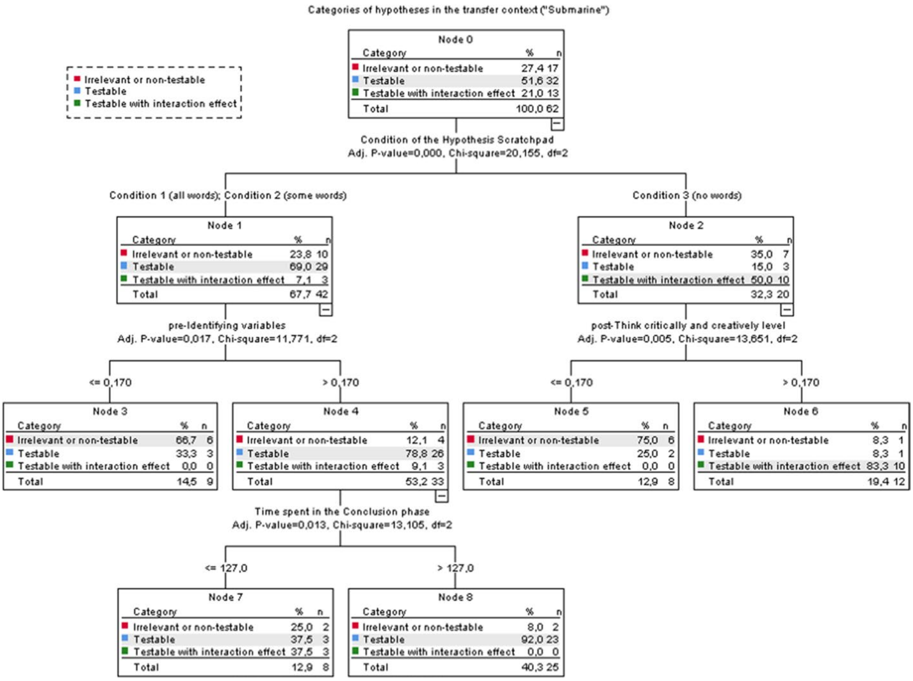

Categories of hypotheses in the transfer context ("Submarine")
■ Irrelevant or non-testable
■ Testable
■ Testable with interaction effect
Node 0
Category % n
Irrelevant or non-testable 27,4 17
Testable 51,6 32
Testable with interaction effect 21,0 13
Total 100,0 62
Condition of the Hypothesis Scratchpad
Adj. P-value=0,000, Chi-square=20,155, df=2
Condition 1 (all words); Condition 2 (some words)
Condition 3 (no words)
Node 1
Category % n
Irrelevant or non-testable 23,8 10
Testable 69,0 29
Testable with interaction effect 7,1 3
Total 67,7 42
Node 2
Category % n
Irrelevant or non-testable 35,0 7
Testable 15,0 3
Testable with interaction effect 50,0 10
Total 32,3 20
pre-Identifying variables
Adj. P-value=0,017, Chi-square=11,771, df=2
post-Think critically and creatively level
Adj. P-value=0,005, Chi-square=13,651, df=2
<= 0,170
> 0,170
<= 0,170
> 0,170
Node 3
Category % n
Irrelevant or non-testable 66,7 6
Testable 33,3 3
Testable with interaction effect 0,0 0
Total 14,5 9
Node 4
Category % n
Irrelevant or non-testable 12,1 4
Testable 78,8 26
Testable with interaction effect 9,1 3
Total 53,2 33
Node 5
Category % n
Irrelevant or non-testable 75,0 6
Testable 25,0 2
Testable with interaction effect 0,0 0
Total 12,9 8
Node 6
Category % n
Irrelevant or non-testable 8,3 1
Testable 8,3 1
Testable with interaction effect 83,3 10
Total 19,4 12
Time spent in the Conclusion phase
Adj. P-value=0,013, Chi-square=13,105, df=2
<= 127,0
> 127,0
Node 7
Category % n
Irrelevant or non-testable 25,0 2
Testable 37,5 3
Testable with interaction effect 37,5 3
Total 12,9 8
Node 8
Category % n
Irrelevant or non-testable 8,0 2
Testable 92,0 23
Testable with interaction effect 0,0 0
Total 40,3 25

Fig. 8 Tree model for categories of hypotheses in the transfer ("submarine") context (domain-specific measure of the targeted skill). Conditions of the Hypothesis Scratchpad, cognitive processes and inquiry skills measured by the pre- and post-test and process variables are shown at each split together with thresholds for partitioning the student sub-sample at each branch (i.e., left and right branches). Each node shows the percentage for each category of hypotheses, number of students (n) and percentage in the total student sample. Percentage of cases correctly classified = 80.6%

contrast, student achievement in the transfer task for Condition 3 was fostered by the cognitive process "Think critically and creatively" as it was recorded in the post-test.

Discussion

The different versions of the HS revealed different outcomes in terms of student achievement in the targeted skill, dependent upon the measure used (see Table 4 for a synopsis of results). These varying outcomes reflect strengths and weaknesses of each version. The fully and partially structured versions of the HS improved the global measure of the targeted skill significantly (TIPSII scale in the pre- and post-test), in contrast to the unstructured version (first research question). On the other hand, the domain-specific measure of the targeted skill revealed that transfer of the ability to identify the interaction effect between the density of the object and fluid was totally absent from the fully structured version of the HS, minimal in the partially structured version and most expressed in the unstructured version (fourth research question). The heterogeneity depicted in Table 4 was observed despite the fact that the global and domain-specific measure were interrelated in two ways, as we presented in our preliminary analysis: First, the global measure was significantly higher for students who identified the interaction effect between the density of the

Table 4 Synopsis of results for the two measures employed for the targeted skill across configurations of the hypothesis scratchpad

	Condition 1 (all words)	Condition 2 (some words)	Condition 3 (no words)
Targeted skill measured by means of TIPSII items in the pre-post test ("Identifying and stating hypotheses"); global measure	Targeted skill improved significantly, on average (see Table 1) Skill improvement was facilitated by learning products in the Investigation Phase (see Fig. 5)	Targeted skill improved significantly, on average (see Table 1) Skill improvement was facilitated by learning products in the Investigation Phase (see Fig. 6)	Targeted skill did not improve significantly, on average (see Table 1) Skill improvement was facilitated by learning products in the Investigation Phase (see Fig. 7) Students overproduced learning products (see Table 2)
Targeted skill measured by means of categories of hypotheses classified by the domain-dependent rubric; domain-specific measure	No student in this condition identified the interaction effect between the density of object and fluid in the transfer context (see Table 3) Generating testable hypotheses (without the interaction effect) was fostered by the inquiry skill "Identifying variables" as recorded already in the pre-test (see Fig. 8)	A small percentage of students in this condition identified the interaction effect between the density of object and fluid in the transfer context (see Table 3) Generating testable hypotheses (without the interaction effect) was fostered by the inquiry skill "Identifying variables" as recorded already in the pre-test (see Fig. 8) Student progression from the intervention to the transfer context was highest among conditions (derived from Table 3; see subsection "Research question 4: How do the three configurations of the Hypothesis Scratchpad differ in their effect on the transfer task, as assessed by the domain-specific measure of the targeted skill?", first paragraph)	Half of the students in this condition identified the interaction effect between the density of object and fluid in the transfer context (see Table 3) Generating testable hypotheses with the interaction effect was fostered by the cognitive process "Think critically and creatively" in the post-test (see Fig. 8)

object and fluid in the intervention context (domain-specific measure); second, improvement in the global measure was significantly higher for students who progressed in the categories of hypotheses they generated when comparing their hypotheses in the intervention context with those generated in the transfer context (domain-specific measure).

An interpretation of the above heterogeneity may be that the full structure may allow students to cultivate an elaborated syntax for generating testable hypotheses (see van Joolingen & de Jong, 1991), and this may be reflected on the improvement of the global measure of the targeted skill (TIPSII scale). However, this same version of the HS may not challenge students enough, up to the point to give them the opportunity to identify the interaction effect between density of the object and fluid (domain-specific measure of the targeted skill) to address the transfer task effectively. Research on worked examples showcased how the strength of fully structured solutions in offering guidance to students may backfire and turn into a major obstacle for transfer, anytime students concentrate on surface features of the problem at hand without processing the task thoroughly to produce a schema, which they can employ in new learning contexts (Eiriksdottir & Catrambone, 2011; Margulieux & Catrambone, 2016). For full and partial structure, the basic entry skill of identifying variables as recorded in the pre-test was enough for students to produce testable hypotheses in the transfer context. In the unstructured condition, instead, transfer was mediated by the cognitive process "Think critically and creatively" recorded after students had exited the intervention context. Half of students in this condition had post-test scores higher than the threshold value indicated by the tree model in this cognitive process and these students were able to detect the interaction effect between the density of object and fluid in the transfer context. This finding implies that the unstructured version of the HS may catalyze transfer, provided that students would be able to employ acquired knowledge and skills in order to filter information in the transfer context and detect the interaction effect as the deep structure of the domain (i.e., "Think critically and creatively").

Taken together, our diverse findings imply that a hypothesis being testable does not secure for effective cognitive processing as far as the deep structure of a domain is concerned. Global measures of the targeted skill may capture testability, but may fail in accounting for schema construction and transfer. Concerning schema construction, a domain-specific measure of the targeted skill should be required. Taking each version of the HS separately, the strength of the fully and partially structured versions of the HS seemed to rely on revealing the testability of hypotheses as indicated by the global measure of the targeted skill. On the other hand, the partially structured and unstructured versions of the HS favored student performance as assessed by the domain-specific measure. The partially structured version was distinguished in terms of highest progression among conditions in the transition from the intervention context to the transfer context (domain-specific measure of the targeted skill). It should be highlighted that the partially structured solution revealed strengths for both measures of the targeted skill (global and domain-specific), which may be attributed to its resemblance to completion problems (partially worked examples). Specifically, partial structure involved desirable features from both structuring the learning task (e.g., partial structure offering guidance to let students initiate the task of hypothesis formulation) and, at the same time, challenging students to complete the task (letting students elaborate on the missing parts of the solution) (see Baars et al., 2013, for a relevant discussion). We recommend that future research should investigate partially worked examples with reference to the distinction between structuring versus "problematizing" student inquiry, as it has been exemplified for software scaffolds (Reiser, 2004). Reiser (2004) contrasted structuring to problematizing: Whereas structuring is required for simplifying open-ended tasks for students, problematizing renders learning trajectories more demanding for students, for instance, by initiating reflection processes and directing student attention to aspects which would, otherwise, remain unaccounted for. Indeed, partially worked examples may strike a delicate balance between different, and, at times, contradictory demands in pedagogical design and instruction, such as managing task complexity (often decreased through structuring but locally increased by "problematizing" to allow for deeper student engagement in learning tasks) and initiative to be undertaken by students in their learning paths (usually withdrawn by structuring to narrow down the options available for students but promoted by "problematizing" student work) (Reiser, 2004, p. 296; Mulder et al., 2016, pp. 505, 507; Xenofontos et al., 2020).

With regard to the unstructured version, schema construction for transfer seems to have been its strength, as indicated for the domain-specific measure of the targeted skill, but this was valid for half of the students in this condition, only. In this case, schema construction for transfer seems to have been conditional upon the development of advanced cognitive processes after students went through the intervention context ("Think critically and creatively"). The unstructured version of the HS was also found to outperform the fully and partially structured conditions in the number of "smart" observations with the interaction effect between the density of the object and fluid (second research question). Other products of learning activities among process variables revealed an analogous trend for students in the unstructured condition, however, these latter cases were not significant (e.g., "smart" trials in the virtual laboratory in the Hypothesis Phase, based on the VOTAT heuristic). Previous research highlighted that fully structured solutions triggered lower student performance than less structured ones (Baars et al., 2013). This was attributed to decreased confidence in student self-efficacy in less structured solutions, which led to overproduction as a compensatory counteraction. It may be that an analogous effect influenced our findings, which needs to be examined by future research. Products of learning activities, furthermore, featured as the most crucial process variables for improving the global measure for the targeted skill across all conditions (third research question), which corroborates the need for an enhanced research focus on the products delivered by students whole enacting learning tasks (see in this regard Hovardas, 2016). We should further note that time-on-task was not linearly connected to going through the learning activity sequence and the delivery of products of learning activities (e.g., Karweit & Slavin, 1982; Slavin, 2014), which is another aspect to be examined by future research.

An additional implication of our results relates to epistemological concerns in addressing deep underlying principles within each different domain and transfer effects. It may be that hypothesis testing should follow after a cycle of inquiry in the domain, which would allow students to explore core assumptions and underlying principles of the studied phenomena. It has been proposed, for example, that inductive synthesis (i.e., seeking explanations that emerge from evidence) rather than hypothetico-deductive analysis (i.e., testing and revising hypotheses) may be more suited to reflect the targeted deep underlying principles (Shemwell et al., 2015). Such an arrangement is in line with the pathways highlighted by Pedaste et al. (2015b) for inquiry-based learning (i.e., exploratory learning trajectory; experimentation learning trajectory). Namely, an exploratory learning trajectory would introduce students to the domain and let them become familiar with the main constituent variables. This stage would not need to include generation of hypotheses, but it could be based on questioning (exploratory learning trajectory; open-ended questions, data-driven approach). After this introductory cycle, students would be better able to engage in the formulation of hypothesis and experimentation (experimentation learning trajectory; theory-based, hypothesis-driven approach). Future research with subsequent cycles of inquiry should take into account such concerns. As long as the skill of hypothesis generation and the transfer of this skill remain indispensable for science education, a single learning context will always prove inadequate to address these related learning and instruction challenges.

Limitations of the study and suggestions for future research

The rather small sample size has been a major limitation of the study, especially in relation to issues of power for the tree models we computed. Random assignment of conditions per classes and not individual students was another limitation. Although this set-up allowed us to avoid any significant differences between conditions in cognitive processes and inquiry skills prior to the educational intervention, future research needs to examine the different conditions of the HS with larger samples and conditions assigned to individual students. Furthermore, our quasi-experimental design with one class per condition, only, cannot allow for ruling out any confoundation. Despite the fact that all conditions were taught by the same teacher, there could be several effects at the class-level which may have influenced our results beyond our control. The short duration of the study, moreover, may have compromised either transfer or the development of additional differences between conditions and effects (see, for instance, Sweller et al., 1998), and therefore, future research should allow students interact with the HS for a longer time frame, and through more learning contexts. Such an arrangement would also allow for withdrawing support (fade out) in line with skill improvement (de Jong, 2006b; Großmann & Wilde, 2019; Kalyuga, 2007) and dynamically adjusting level/type of scaffolding needs to individual student performance (Molenaar & Roda, 2008; see also Kao et al., 2017, for an elaboration on the need to customize scaffolds for addressing instructional goals). An increased variability of contexts would also add to the generalizability of our findings beyond the domain of relative density. In addition, future research should employ more direct and detailed measures of time-on-task than log file data and operationalize more diverse learning products so as to monitor the effects of process variables on the targeted skill and transfer.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11251-022-09580-x.

Acknowledgements The present study was undertaken within the frame of the research project "Go-Lab: Global Online Science Labs for Inquiry Learning at School", funded by the European Community (Grant Agreement No. 317601; Information and Communication Technologies (ICT) theme; 7th Framework Programme for R&D). We are grateful to our colleagues Ellen T. Wassink-Kamp, Casper H. W. de Jong, Margus Pedaste, Mario Mäeots, Leo Siiman, Effie Law, Matthias Heintz, and Rob Edlin White for their help in developing the rubric to score hypotheses.

References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Addison Wesley Longman.

Arnold, J. C., Kremer, K., & Mayer, J. (2014). Understanding students' experiments—what kind of support do they need in inquiry tasks? *International Journal of Science Education, 36*, 2719–2749. https://doi.org/10.1080/09500693.2014.930209

Baars, M., Visser, S., van Gog, T., Bruin, A. D., & Paas, F. (2013). Completion of partially worked-out examples as a generation strategy for improving monitoring accuracy. *Contemporary Educational Psychology, 38*, 395–406. https://doi.org/10.1016/j.cedpsych.2013.09.001

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Belenky, D. M., & Schalk, L. (2014). The effects of idealized and grounded materials on learning, transfer, and interest: An organizing framework for categorizing external knowledge representations. *Educational Psychology Review, 26*, 27–50. https://doi.org/10.1007/s10648-014-9251-9

Bell, T., Urhahne, D., Schanze, S., & Ploetzner, R. (2010). Collaborative inquiry learning: Models, tools and challenges. *International Journal of Science Education, 32*, 349–377. https://doi.org/10.1080/09500690802582241

Bloom, B. S. (1956). *Taxonomy of educational objectives. Handbook I: The cognitive domain*. David McKay.

Burns, J. C., Okey, J. R., & Wise, K. C. (1985). Development of an integrated process skill test: TIPS II. *Journal of Research in Science Teaching, 22*, 169–177. https://doi.org/10.1002/tea.3660220208

Chang, K. E., Chen, Y. L., Lin, H. Y., & Sung, Y. T. (2008). Effects of learning support in simulation-based physics learning. *Computers & Education, 51*, 1486–1498. https://doi.org/10.1016/j.compedu.2008.01.007

Chen, J., Wang, M., Grotzer, T. A., & Dede, C. (2018). Using a three-dimensional thinking graph to support inquiry learning. *Journal of Research in Science Teaching, 55*, 1239–1263. https://doi.org/10.1002/tea.21450

Chi, M. T. H., & VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist, 47*, 177–188. https://doi.org/10.1080/00461520.2012.695709

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). Routledge.

de Jong, T. (2006a). Scaffolds for scientific discovery learning. In J. Elen & R. E. Clark (Eds.), *Handling complexity in learning environments: Theory and research* (pp. 107–128). London: Elsevier.

de Jong, T. (2006b). Computer simulations – Technological advances in inquiry learning. *Science*, *312*, 532–533. https://doi.org/10.1126/science.1127750

de Jong, T. (Ed.). (2014). Preliminary inquiry classroom scenarios and guidelines. D1.3. Go-Lab Project (Global Online Science Labs for Inquiry Learning at School).

de Jong, T., Gillet, D., Rodríguez-Triana, M. J., Hovardas, T., Dikke, D., Doran, R., Dziabenko, O., Koslowsky, J., Korventausta, M., Law, E., Pedaste, M., Tasiopoulou, E., Vidal, G., & Zacharia, Z. C. (2021). Understanding teacher design practices for digital inquiry–based science learning: The case of Go-Lab. *Educational Technology Research & Development, 69*, 417–444. https://doi.org/10.1007/s11423-020-09904-z

de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*, 179–202. https://doi.org/10.3102/00346543068002179

de Jong, T., Sotiriou, S., & Gillet, D. (2014). Innovations in STEM education: the Go-Lab federation of online labs. *Smart Learning Environments*, *1*, 1–16. https://doi.org/10.1186/s40561-014-0003-6

Efstathiou, C., Hovardas, T., Xenofontos, N., Zacharia, Z., de Jong, T., Anjewierden, A., & van Riesen S. A. N. (2018). Providing guidance in virtual lab experimentation: The case of an experiment design tool. *Educational Technology Research & Development*, *66*, 767–791. https://doi.org/10.1007/s11423-018-9576-z

Gijlers, H., & de Jong, T. (2005). The relation between prior knowledge and students' collaborative discovery learning processes. *Journal of Research in Science Teaching*, *42*, 264–282. https://doi.org/10.1002/tea.20056

Gijlers, H., & de Jong, T. (2009). Sharing and confronting propositions in collaborative inquiry learning. *Cognition and Instruction*, *27*, 239–268. https://doi.org/10.1080/07370000903014352

Hovardas, T. (2016). A learning progression should address regression: Insights from developing non-linear reasoning in ecology. *Journal of Research in Science Teaching*, *53*, 1447–1470. https://doi.org/10.1002/tea.21330

Eiriksdottir, E., & Catrambone, R. (2011). Procedural instructions, principles, and examples: How to structure instructions for procedural tasks to enhance performance, learning, and transfer. *Human Factors, 53*, 749–770. https://doi.org/10.1177/0018720811419154

Großmann, N., & Wilde, M. (2019). Experimentation in biology lessons: Guided discovery through incremental scaffolds. *International Journal of Science Education, 41*, 759–781. https://doi.org/10.1080/09500693.2019.1579392

Hmelo-Silver, S. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*, 99–107. https://doi.org/10.1080/00461520701263368

Hsin, C.-T., & Wu, H.-K. (2011). Using scaffolding strategies to promote young children's scientific understandings of floating and sinking. *Journal of Science Education and Technology, 20*, 656–666. https://doi.org/10.1007/s10956-011-9310-7

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509–539. https://doi.org/10.1007/s10648-007-9054-3

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. *Science, 320*, 454–455. https://doi.org/10.1126/science.1154659

Kao, G. Y. M., Chiang, C. H., & Sun, C. T. (2017). Customizing scaffolds for game-based learning in physics: Impacts on knowledge acquisition and game design creativity. *Computers & Education, 113*, 294–312. https://doi.org/10.1016/j.compedu.2017.05.022

Karweit, N., & Slavin, R. E. (1982). Time-on-task: Issues of timing, sampling, and definition. *Journal of Educational Psychology, 74*, 844–851. https://doi.org/10.1037/0022-0663.74.6.844

Kim, J. H., & Pedersen, S. (2011). Advancing young adolescents' hypothesis- development performance in a computer-supported and problem-based learning environment. *Computers & Education, 57*, 1780–1789. https://doi.org/10.1016/j.compedu.2011.03.014

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75–86. https://doi.org/10.1207/s15326985ep4102_1

Klahr, D. (2005). A framework for cognitive studies and technology. In M. Gorman, R. D. Tweney, D. C. Gooding, & A. P. Kincannon (Eds.), *Scientific and technological thinking* (pp. 81–95). Lawrence Erlbaum.

Koksal, E. A., & Berberoglou, G. (2014). The effect of guided inquiry instruction on 6th grade Turkish students' achievement, science process skills, and attitudes toward science. *International Journal of Science Education, 36*, 66–78. https://doi.org/10.1080/09500693.2012.721942

Loverude, M. E., Kautz, C. H., & Heron, P. R. L. (2003). Helping students develop an understanding of Archimedes' principle. I. Research on student understanding. *American Journal of Physics, 71*, 1178–1187. https://doi.org/10.1119/1.1607335

Margulieux, L. E., & Catrambone, R. (2016). Improving problem solving with subgoal labels in expository text and worked examples. *Learning and Instruction, 42*, 58–71. https://doi.org/10.1016/j.learninstruc.2015.12.002

Meindertsma, H. B., van Dijk, M. W. G., Steenbeek, H. W., & van Geert, P. L. C. (2014). Stabilty and variability in young children's understanding of floating and sinking during one single-task session. *Mind, Brain, and Education, 8*, 149–158. https://doi.org/10.1111/mbe.12049

Molenaar, I., & Roda, C. (2008). Attention management for dynamic and adaptive scaffolding. *Pragmatics & Cognition, 16*, 224–271. https://doi.org/10.1075/pc.16.2.04mol

Mulder, Y. G., Bollen, L., de Jong, T., & Lazonder, A. W. (2016). Scaffolding learning by modelling: The effects of partially worked-out models. *Journal of Research in Science Teaching*, *53*, 502–523. https://doi.org/10.1002/tea.21260

Oh, P. S. (2010). How can teachers help students formulate scientific hypotheses? Some strategies found in abductive inquiry activities of earth science. *International Journal of Science Education, 32*, 541–560. https://doi.org/10.1080/09500690903104457

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*, 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Potvin, P., & Cyr, G. (2017). Toward a durable prevalence of scientific conceptions: Tracking the effects of two interfering misconceptions about buoyancy from preschoolers to science teachers. *Journal of Research in Science Teaching, 54*, 1121–1142. https://doi.org/10.1002/tea.21396

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*, 337–386. https://doi.org/10.1207/s15327809jls1303_4

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences, 13*, 273–304. https://doi.org/10.1207/s15327809jls1303_2

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology, 103*, 759–775. https://doi.org/10.1037/a0025140

Shemwell, J. T., Chase, C. C., & Schwartz, D. L. (2015). Seeking the general explanation: A test of inductive activities for learning and transfer. *Journal of Research in Science Teaching, 52*, 58–83. https://doi.org/10.1002/tea.21185

Slavin, R. E. (2014). *Educational psychology: Theory and practice* (11th ed.). Pearson Education.

Sweller, J., van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296. https://doi.org/10.1023/A:1022193728205

Sweller, J., Kirschner, P. A., & Clark, R. E. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. *Educational Psychologist, 42*, 115–121. https://doi.org/10.1080/00461520701263426

Van Merriënboer, J. J. G. (1990). Strategies for programming instruction in high school: Program completion vs. program generation. *Journal of Educational Computing Research, 6*, 265–285. https://doi.org/10.2190/4NK5-17L7-TWQV-1EHL

van Joolingen, W. R., & de Jong, T. (1991). Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instructional Science*, *20*, 389–404. https://doi.org/10.1007/BF00116355

van Joolingen, W. R., & de Jong, T. (1993). Exploring a domain through a computer simulation: Traversing variable and relation space with the help of a hypothesis scratchpad. In D. Towne, T. de Jong, & H. Spada (Eds.), *Simulation-based experiential learning* (pp. 191–206). (NATO ASI series). Berlin: Springer.

van Joolingen, W. R., & de Jong, T. (1997). An extended dual search space model of learning with computer simulations. *Instructional Science*, *25*, 307–346. https://doi.org/10.1023/A:1002993406499

van Joolingen, W. R., & de Jong, T. (2003). SimQuest: authoring educational simulations. In T. Murray, S. Blessing, & S. Ainsworth (Eds.), *Authoring tools for advanced technology educational software: Toward cost-effective production of adaptive, interactive, and intelligent educational software* (pp. 1–31). Dordrecht: Kluwer Academic Publishers.

van Joolingen, W. R., de Jong, T., Lazonder, A. W., Savelsbergh, E. R., & Manlove, S. (2005). Co-Lab: Research and development of an online learning environment for collaborative scientific discovery learning. *Computers in Human Behavior*, *21*, 671–688. https://doi.org/10.1016/j.chb.2004.10.039

Van Merriënboer, J. J. G., & de Croock, M. B. M. (1992). Strategies for computer-based programming instruction—program completion vs program generation. *Journal of Educational Computing Research, 8*, 365–394. https://doi.org/10.2190/MJDX-9PP4-KFMT-09PM

Xenofontos, N. A., Hovardas, T., Zacharia, Z. C., & de Jong, T. (2020). Inquiry-based learning and retrospective action: Problematizing student work in a computer-supported learning environment. *Journal of Computer Assisted Learning*, *36*, 12-28. https://doi.org/10.1111/jcal.12384

Zacharia, Z. C., Manoli, C., Xenofontos, N., de Jong, T., Pedaste, M., van Riesen, S., Kamp, E., Mäeots, M., Siiman. L., & Tsourlidaki, E. (2015). Identifying potential types of guidance for supporting student inquiry when using virtual and remote labs: A literature review. *Educational Technology Research and Development*, *63*, 257–302. https://doi.org/10.1007/s11423-015-9370-0

Zervas, P. (Ed.). (2013). The Go-Lab inventory and integration of online labs—Labs offered by large scientific organisations. D2.1. Go-Lab Project (Global Online Science Labs for Inquiry Learning at School).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.